# 7.贝叶斯分类器

## 7.1.贝叶斯决策论

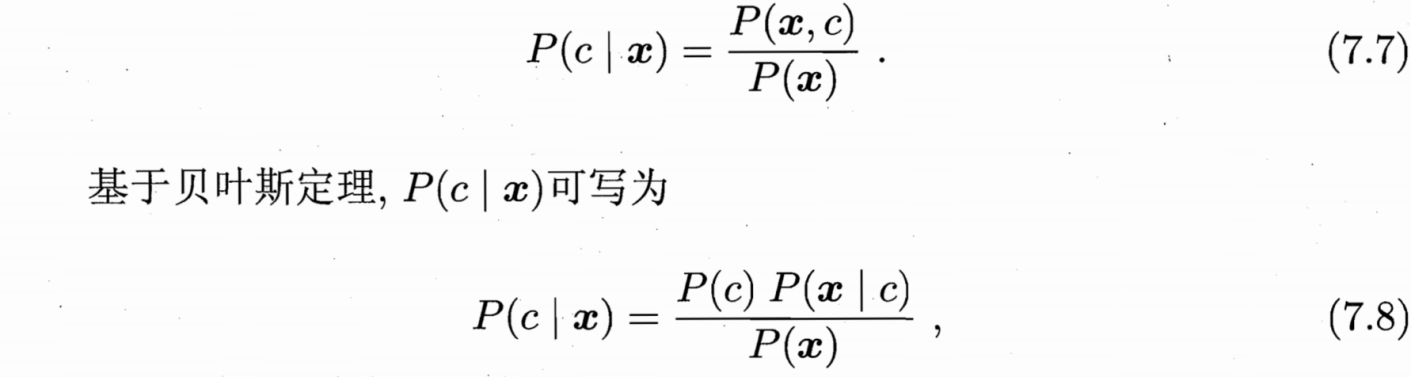
要实现基于有限的训练样本集尽可能准确地估计出后验概率，大体来说，主要有两种策略：

“判别式模型”（discriminative models）：给定，可通过直接建模来预测，前面介绍的决策树、BP神经网络、支持向量机等，都可归入判别式模型的范畴。

“生成式模型”（generative models）：先对联合概率分布建模，然后再由此获得，贝叶斯就是生成模型。

贝叶斯公式

利用联合概率我们可以计算出条件概率



其中，是类“先验”（prior）概率；是样本相对于类标记c的类条件概率（class-conditional probability），或称为“似然”（likelihood）；是用于归一化的“证据”（evidence）因子。对给定样本，证据因子与类标记无关，因此估计的问题就转化为如何基于训练数据来估计先验和似然。

类先验概率表达了样本空间中各类样本所占的比例，根据大数定律，当训练集包含充足的独立同分布样本时，可通过各类样本出现的频率来进行估计.

对类条件概率来说，由于它涉及关于所有属性的联合概率，直接根据样本出现频数来估计会遇到严重的问题。因此这里一般采用极大似然法进行估计。

现举例说明如何利用贝叶斯做决策

假设分别收集北半球和南半球100个人的生日作为样本，1到12月生人的分布情况为：

北半球：3     4     5     7    10    13    14    15    12     8     5     4

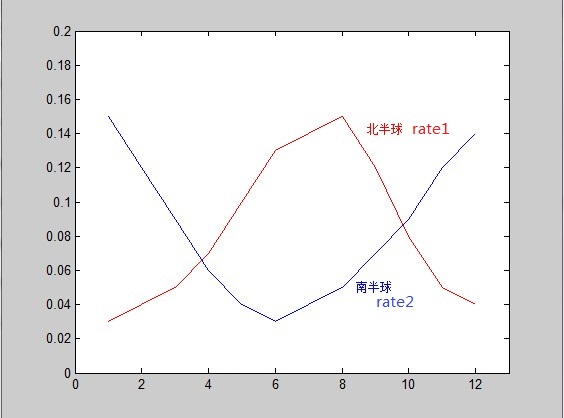
南半球：15    12     9     6     4     3     4     5     7     9    12    14

那么1月到12月生日所占的比率分别为：

北半球：0.03    0.04    0.05    0.07    0.10    0.13    0.14    0.15    0.12    0.08    0.05    0.04

南半球：0.15    0.12    0.09    0.06    0.04    0.03    0.04    0.05    0.07    0.09    0.12    0.14

可以绘制概率密度曲线如下图



设一个人为北半球人事件为，为南半球人事件为，则，由百度得，，令生日为变量，当知道一个人生日为6月，那么他来自北半球还是南半球呢？

首先列出贝叶斯公式



这里，，则由题得

，



所以





所以这个人有97.5%的可能性是来自北半球的，只有2.5%的可能性是来自南半球。

我们可以设定这样的判决规则：

如果，则判决为类，否则为类

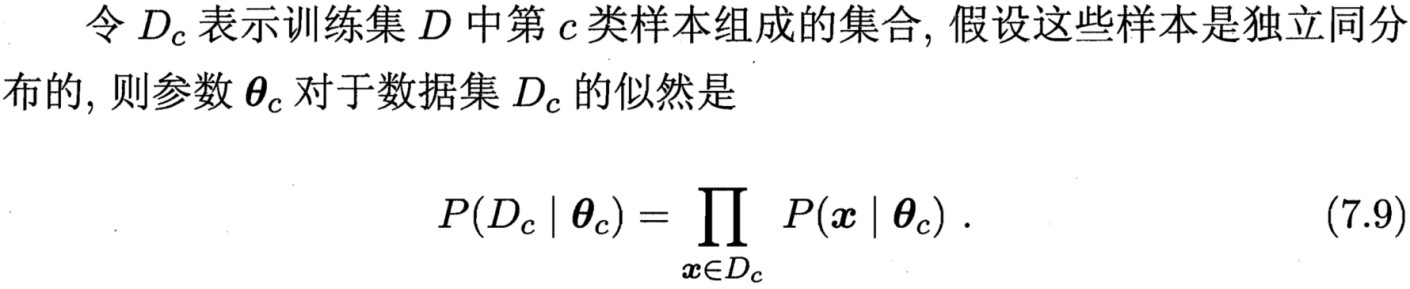
根据上面的规则，判定这个人是来自类，即北半球。

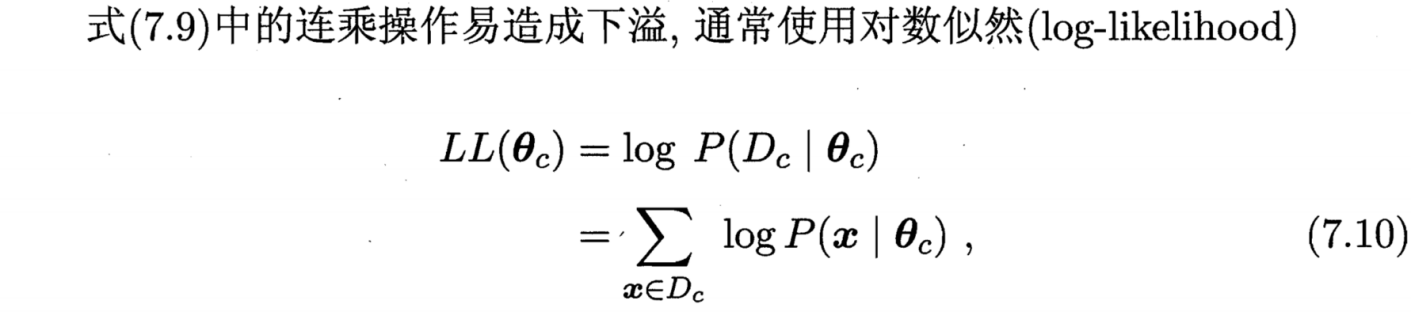
 从上述例子中可以看出，证据因子其实对做出某种判决并不重要，只是用于最后的归一化，因此在计算贝叶斯时通常省略这个项。

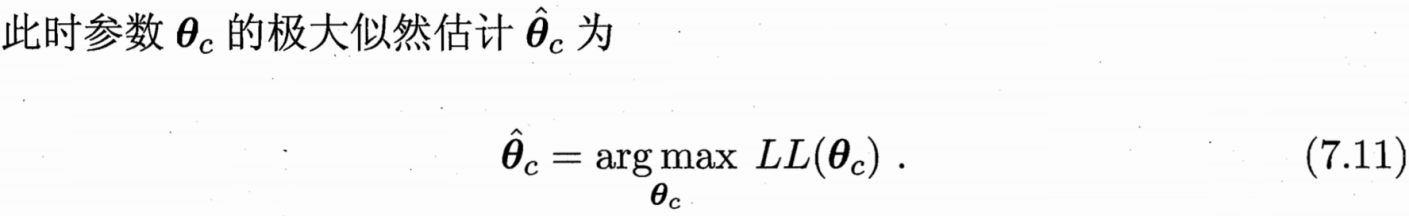
## 7.2.极大似然估计（MLE）

极大似然估计（Maximum Likelihood Estimation，简称MLE），是一种根据数据采样来估计概率分布的经典方法。

极大似然法的核心思想就是：利用已知的样本结果，反推最有可能（最大概率）导致这样结果的参数值。







总结最大似然法估计参数的过程，一般分为以下四个步骤：

\* 1.写出似然函数；

\* 2.对似然函数取对数，并整理；

\* 3.求导数，令偏导数为0，得到似然方程组；

\* 4.解似然方程组，得到所有参数即为所求。

例：现将一枚6面骰子抛掷10次，抛掷出的点数分别为2、3、2、5、4、.6、1、3、4、2，试基于此抛掷结果估计这枚骰子抛掷出各个点数的概率。

解：设这枚骰子抛掷出点数的概率为，根据极大似然估计法可以写出似然函数为



其对数似然函数即为



由于之间满足如下约束



所以此时最大化对数似然函数属于带约束的最优化问题，也即

由拉格朗日乘子法可得拉格拉格朗日函数为



对拉格朗日函数分别关于求偏导，然后令其等于0可得









同理可求得



又因为



所以最终解得

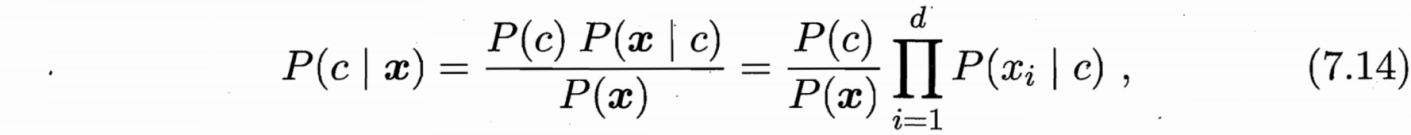
，，，，，

此时抛掷出各个点数的概率值与其频率值相等

## 7.3.朴素贝叶斯分类器

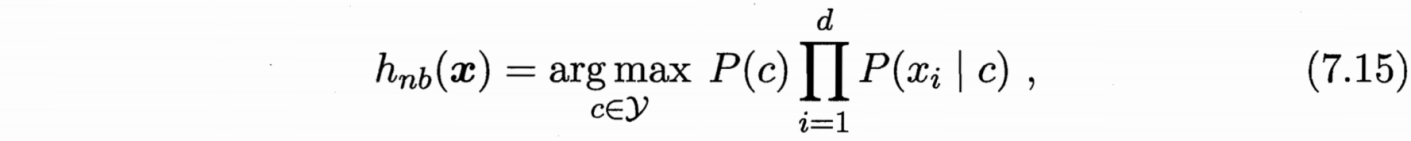
朴素贝叶斯法（naive Bayes）是基于贝叶斯定理与属性条件独立假设的分类方法。

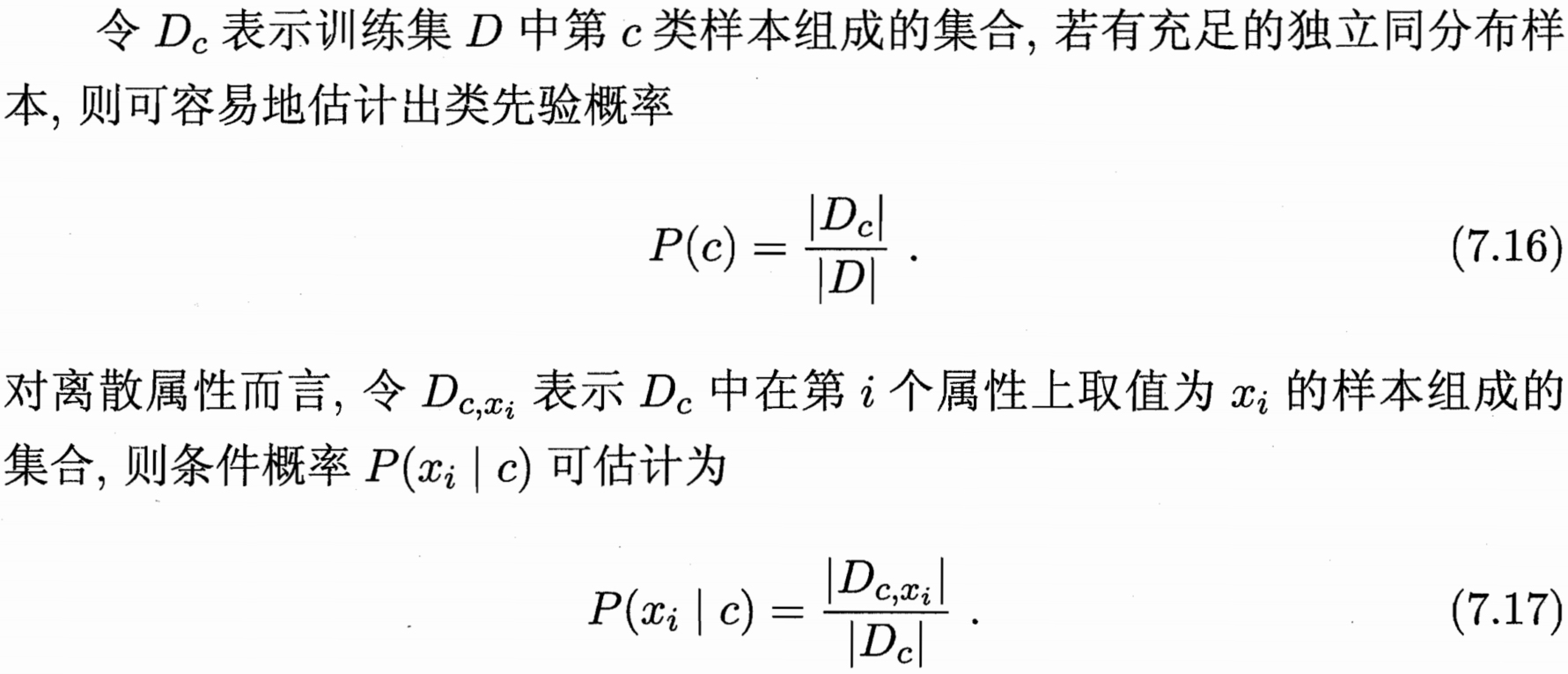
基于属性条件独立性假设，贝叶斯公式可重写为

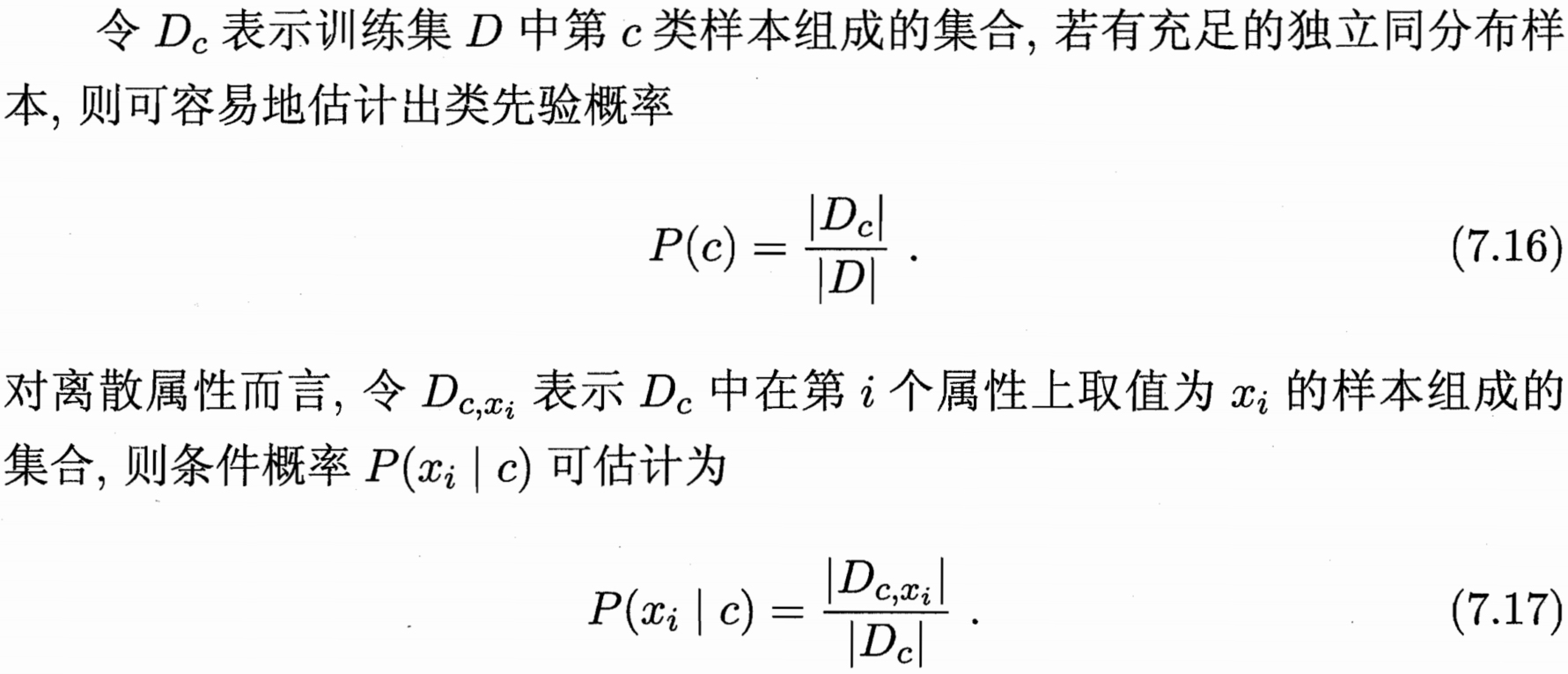


其中为属性数目，为在第个属性上的取值.

由于对所有类别来说相同，因此基于上式的朴素贝叶斯分类器的表达式为



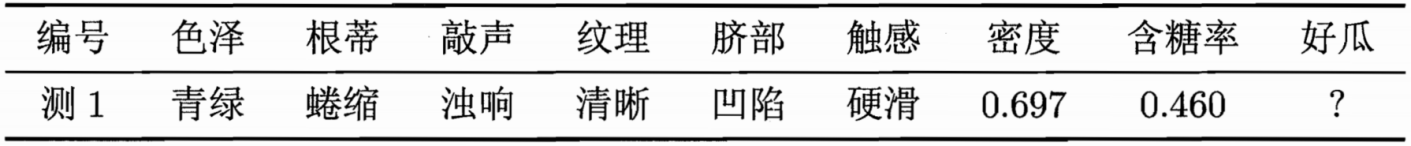


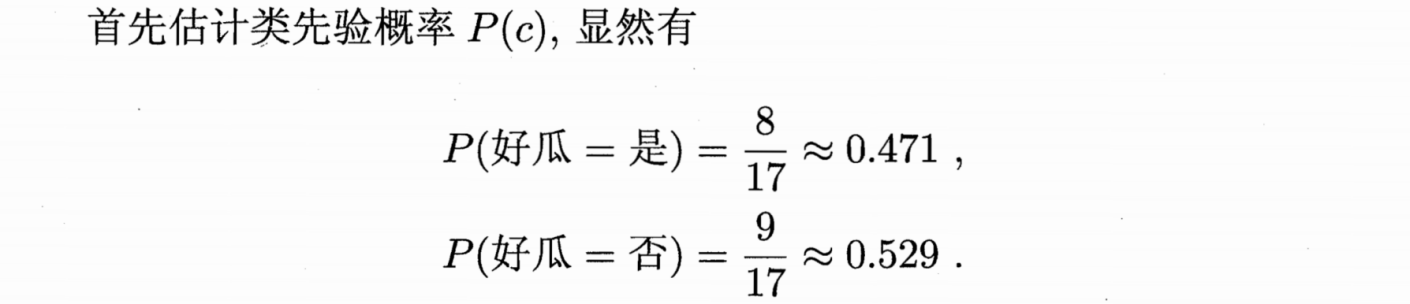


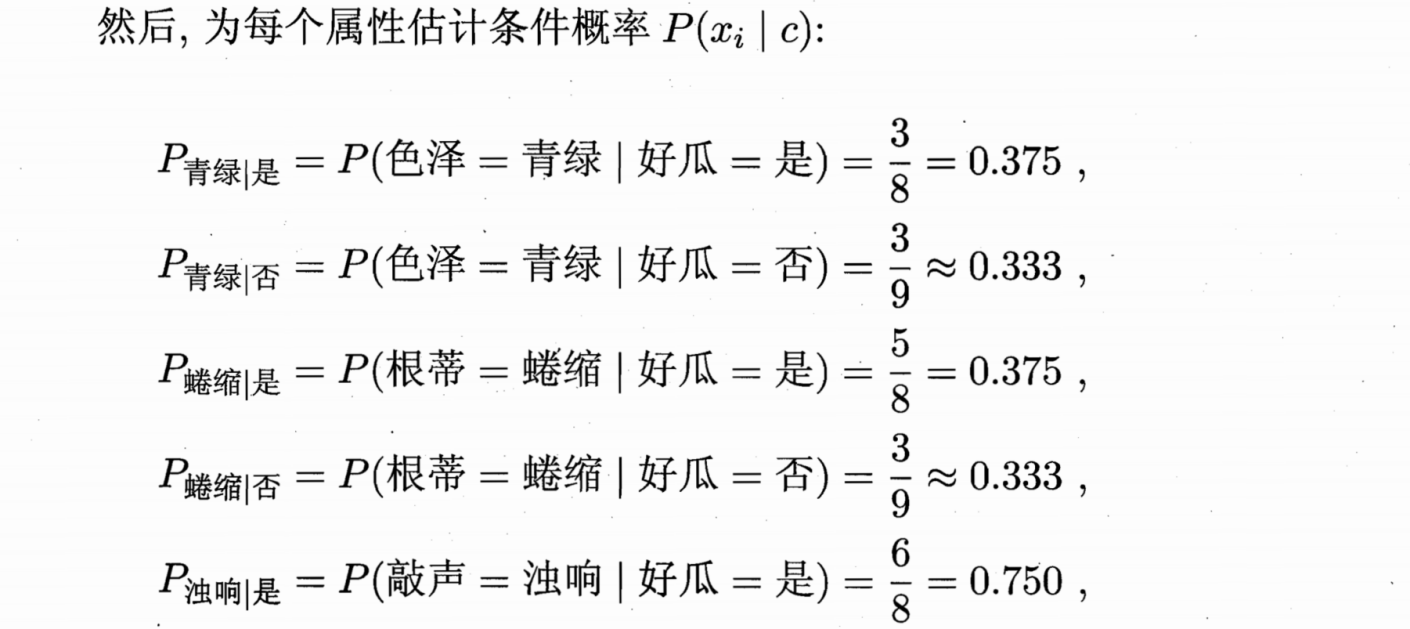
下面我们用西瓜数据集3.0训练一个朴素贝叶斯分类器，对测试例“测1”进行分类：

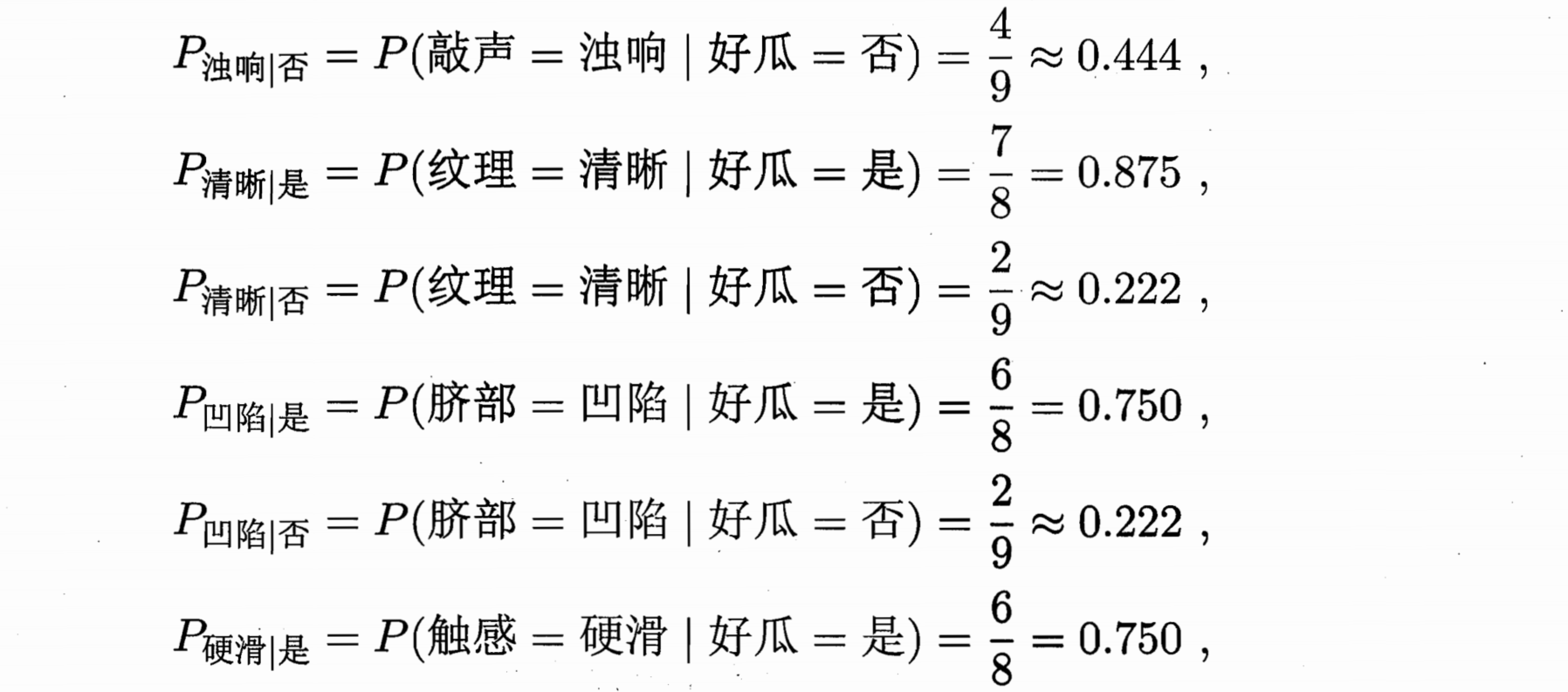


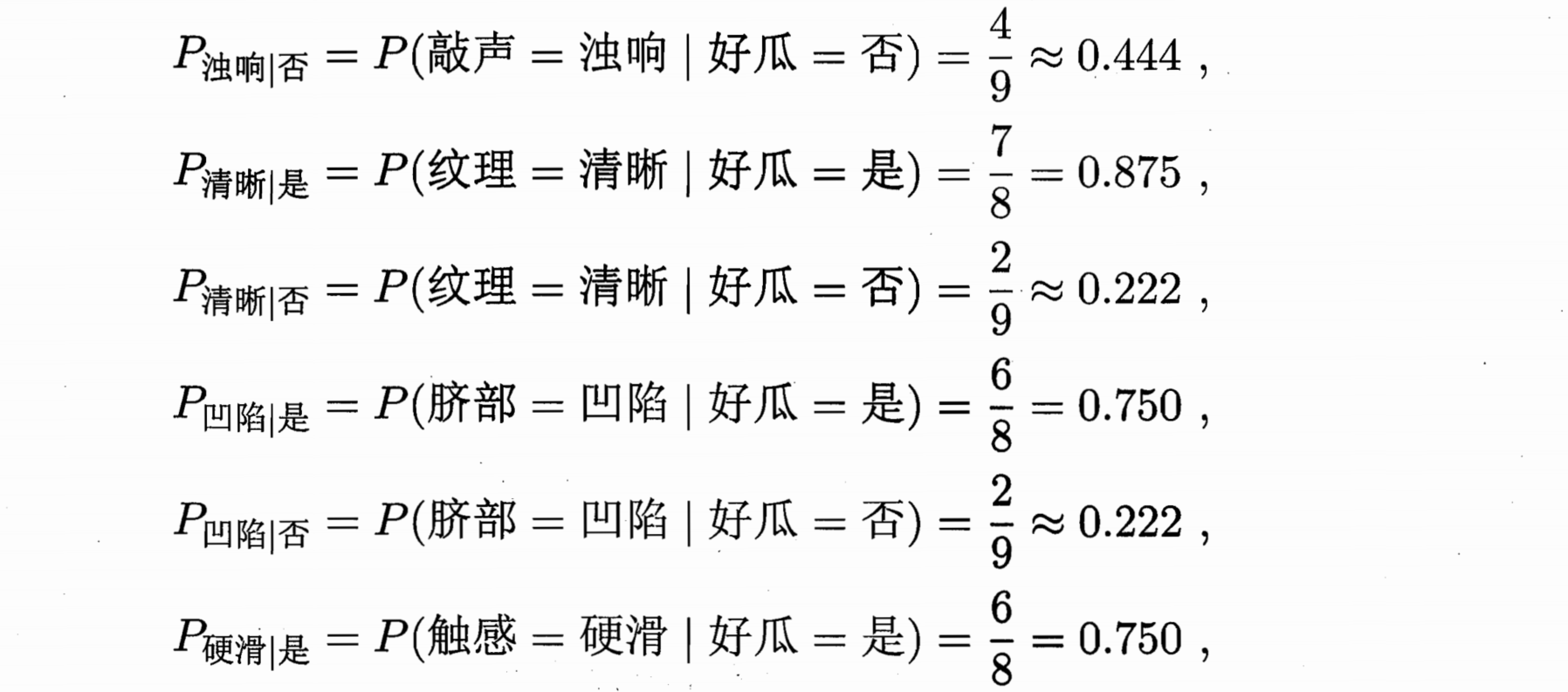


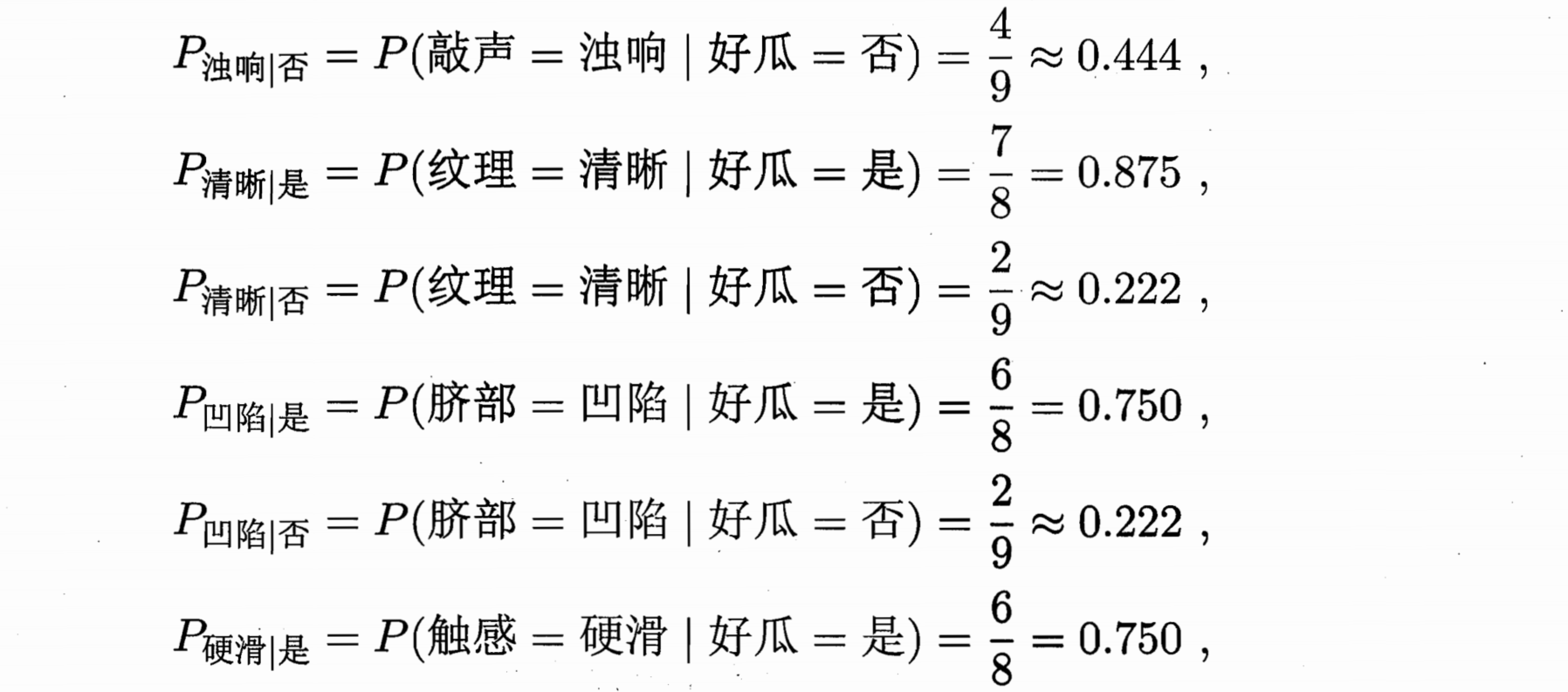


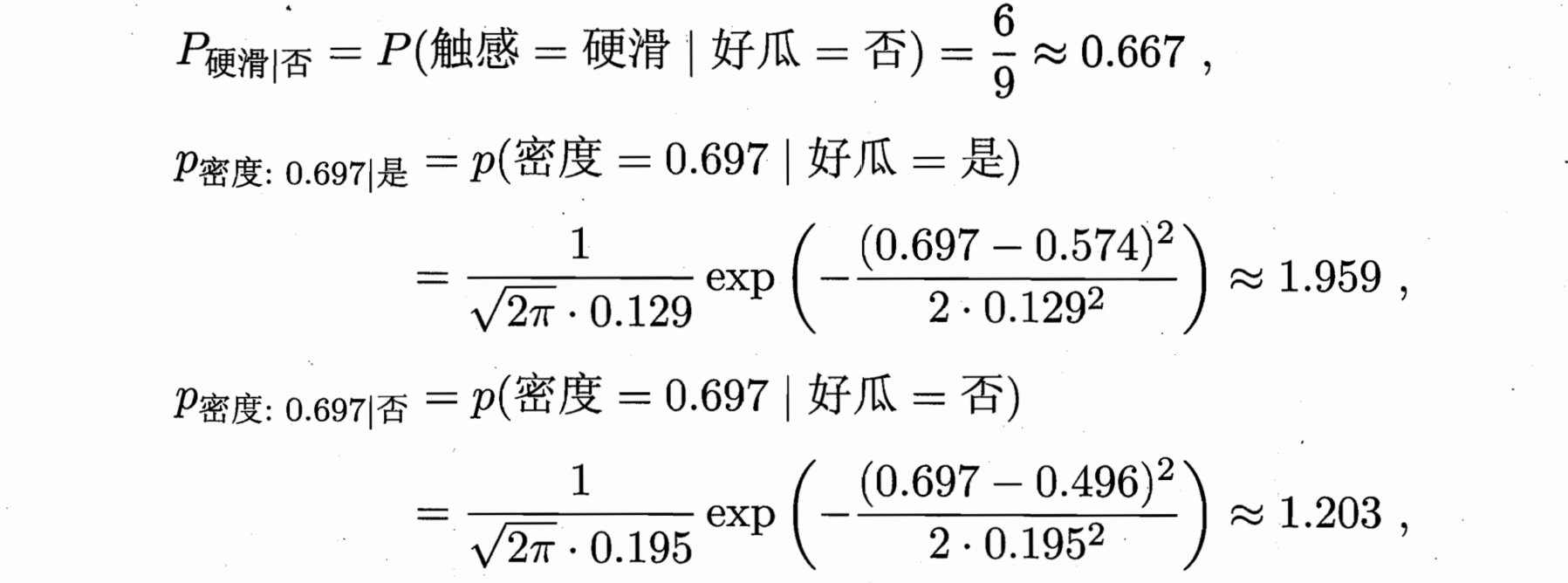




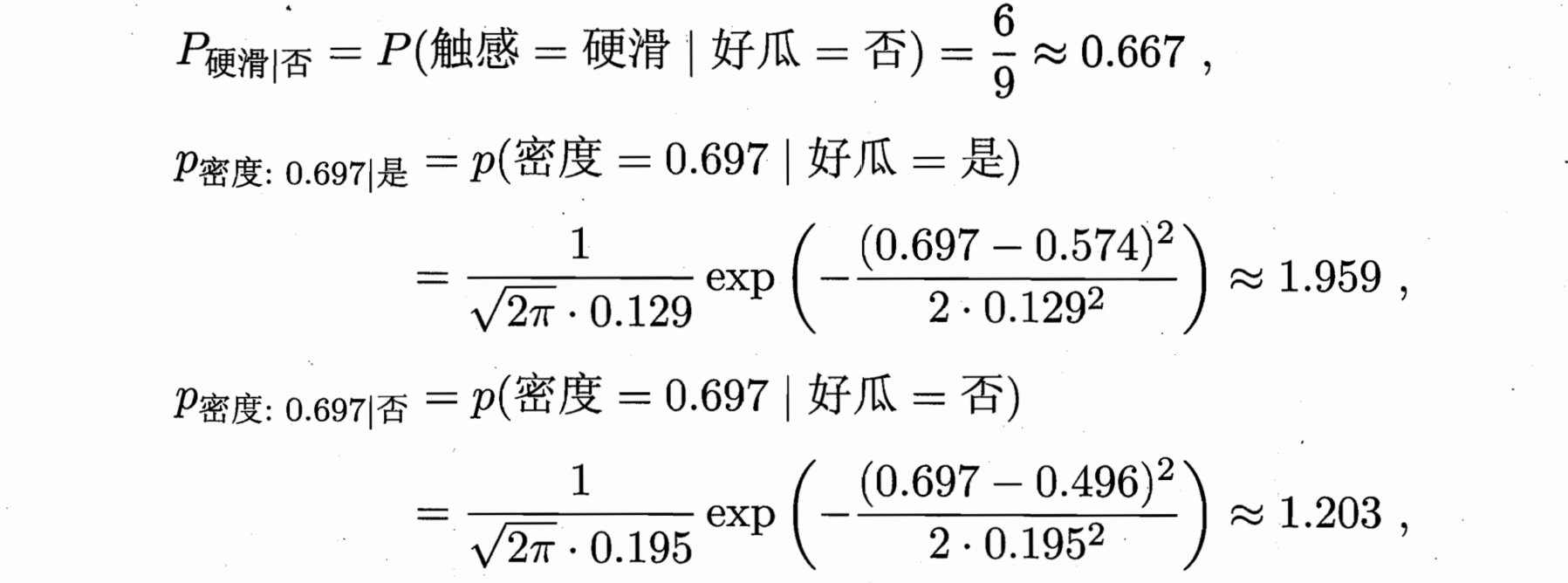




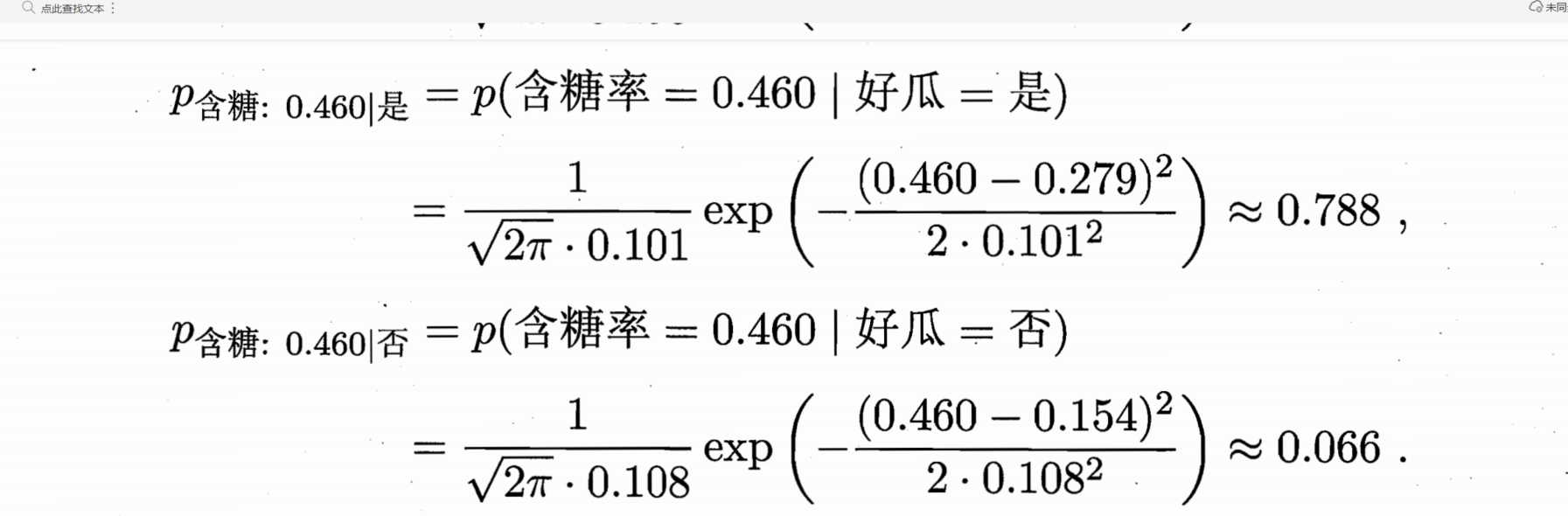




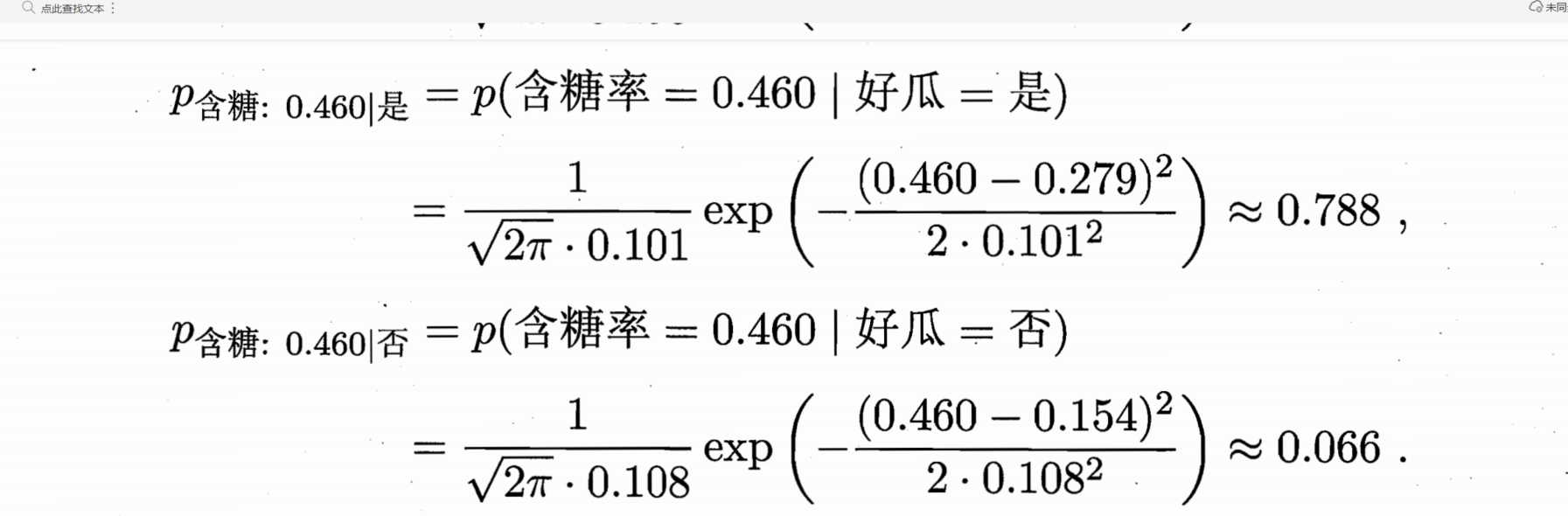
其中，0.574为好瓜的密度均值，为方差



其中，0.496为坏瓜的密度均值，为方差

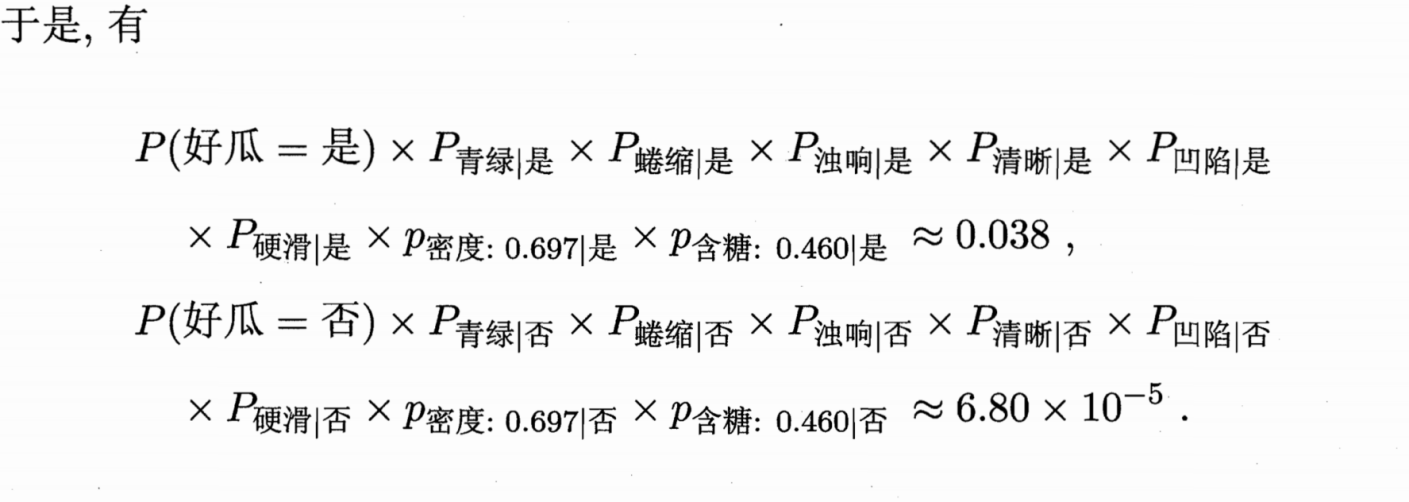


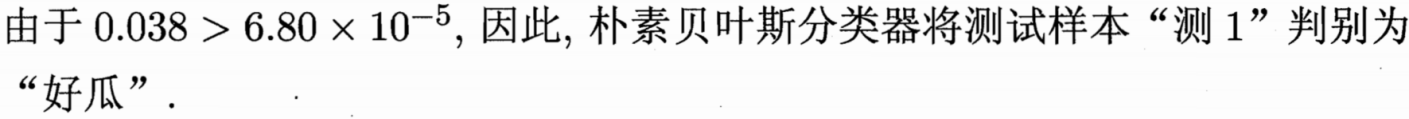
其中，0.297为好瓜的含糖率均值，为方差



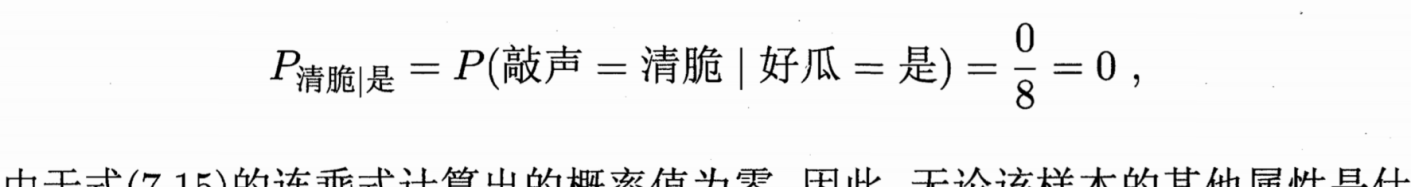
其中，0.154为坏瓜的含糖率均值，为方差

于是有



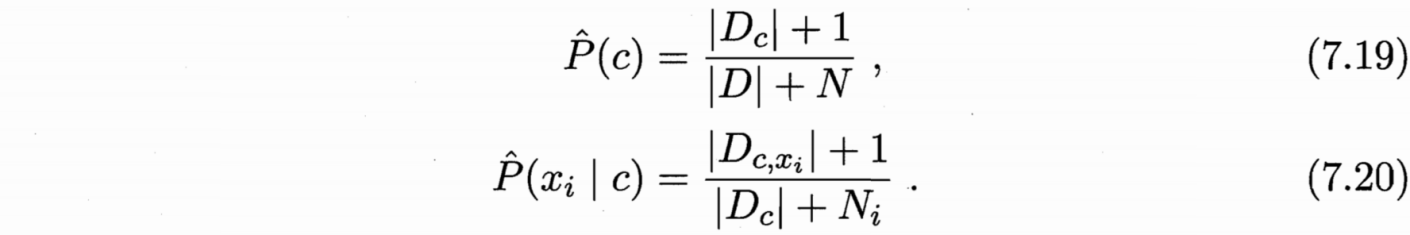


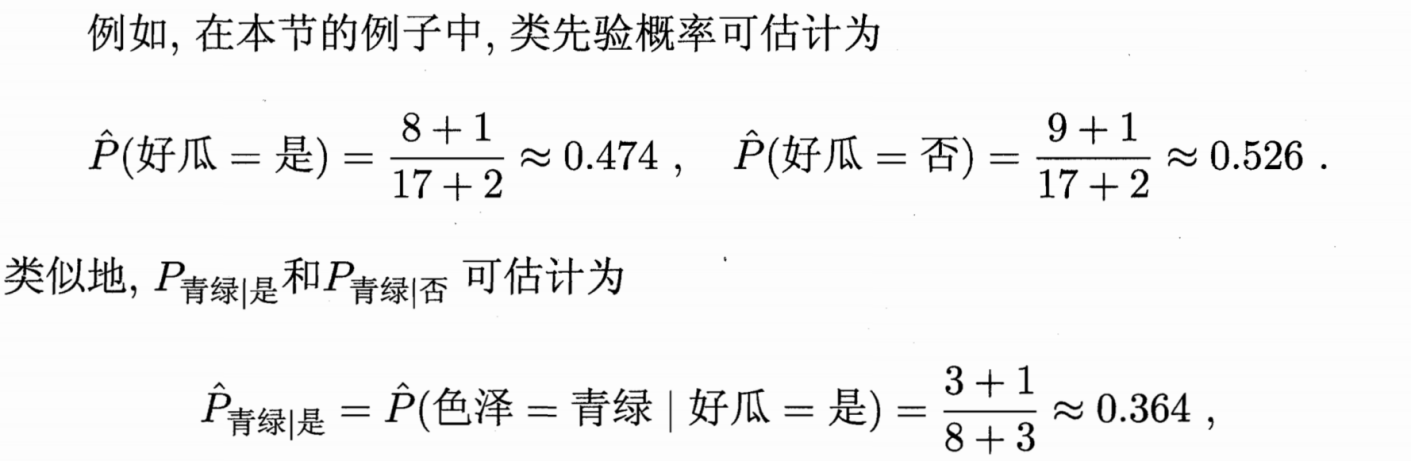
需要注意，若某个属性值在训练集中没有与某个类同时出现过，如

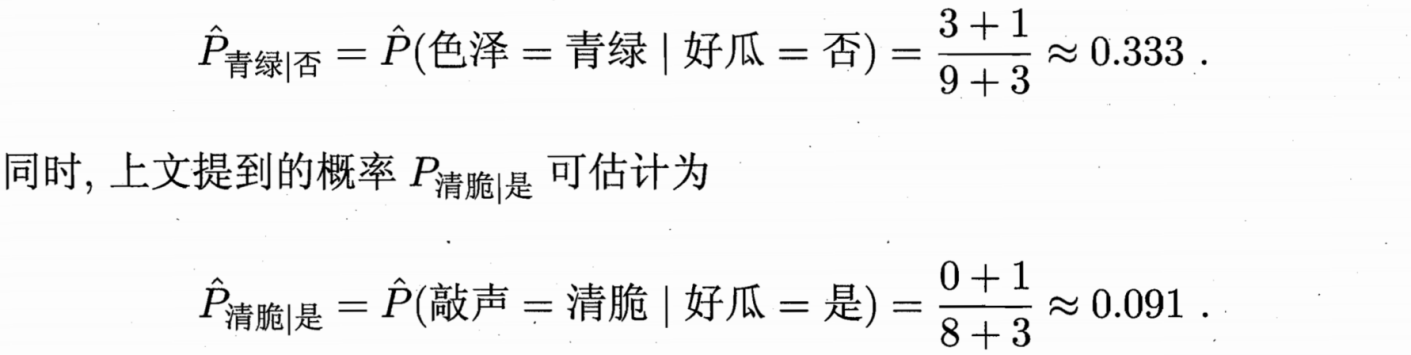


此时，连乘式计算出的概率值为0。

为了避免此问题，在估计概率值时通常要进行“平滑”（smoothing），常用“拉普拉斯修正”（Laplacian correction）。具体来说，令表示训练集中可能的类别数，表示第个属性可能的取值数，则式（7.16）和（7.17）分别修正为







显然，此方法避免了因数据集样本不充分而导致概率值估计为0的问题。常用的平滑方式还有古德-图灵（Good-Turing）估计。

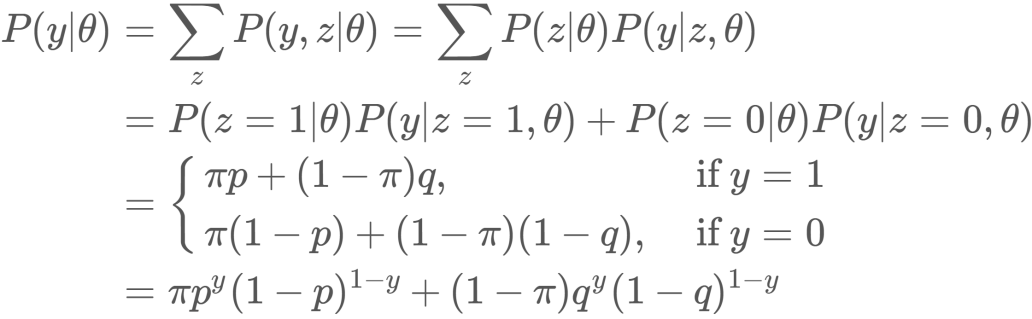
## 7.4.EM算法

概率模型有时既含有观测变量，又含有隐变量或者潜在变量。如果概率模型的变量都是观测变量，那么给定数据，可以直接用极大似然估计法，或者贝叶斯估计法估计模型参数。但是，当模型含有隐变量时，就不能简单地使用这些估计方法。  
 EM算法就是含有隐变量的概率模型参数的极大似然估计法。

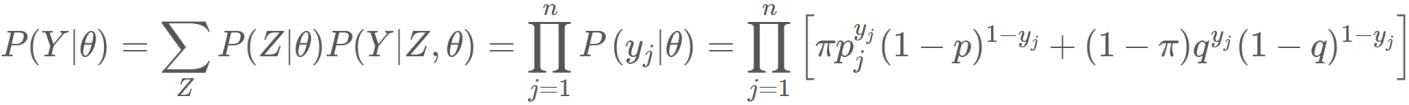
首先介绍一个使用EMEM算法的例子：三硬币模型

假设有3枚硬币，分别记作A，B，C。这些硬币正面出现的概率分别是。试验如下：先掷硬币A，根据其结果选出硬币B或硬币C，正面选硬币B，反面选硬币C；然后掷选出的硬币，出现正面记作1，出现反面记作0；独立地重复n次实验（这里，n=10），观测结果如下：  
 1，1，0，1，0，0，1，0，1，1  
假设只能观测到掷硬币的结果，不能观测掷硬币的过程。问如何估计三硬币正面出现的概率，即三硬币模型的参数。

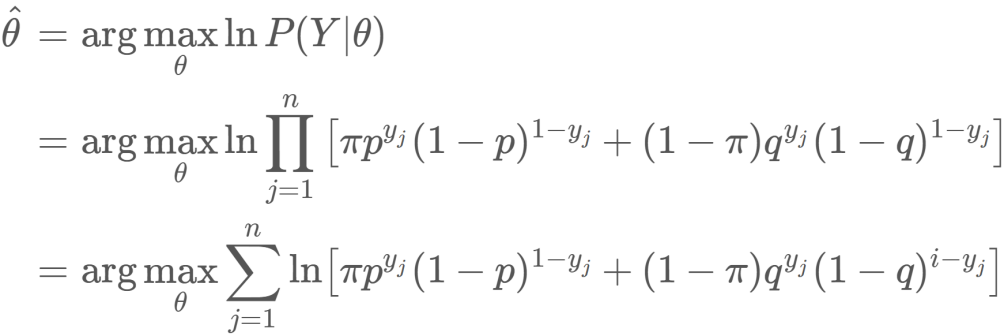
解：对每一次试验可以进行如下建模



这里，随机变量是观测变量，表示一次试验观测的结果1或0；随机变量是隐变量，表示未观测到的掷硬币A的结果；是模型参数。将观测数据表示为，未观测数据表示为，则观测数据的似然函数为：



考虑求模型参数的极大似然估计，即使用对数似然函数来进行参数估计可得：



上式没有解析解，也就是没办法直接解出恰好等于某个常数，只能用送代的方法来进行求解。EM算法就是可以用于求解这个问题的一种迭代算法。

在讲解EM算法前，首先引入**Jensen（琴生）不等式**

若是凸函数，则：



其中，，代表和之间任意的点

将上式中的t推广到n个同样成立，也即：



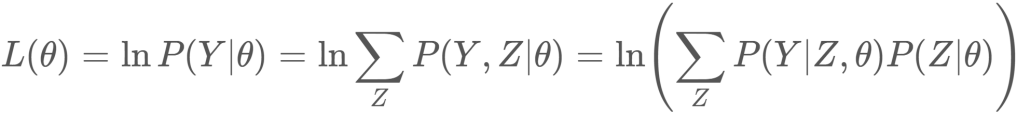
其中，，。在概率论中常以以下形式出现：



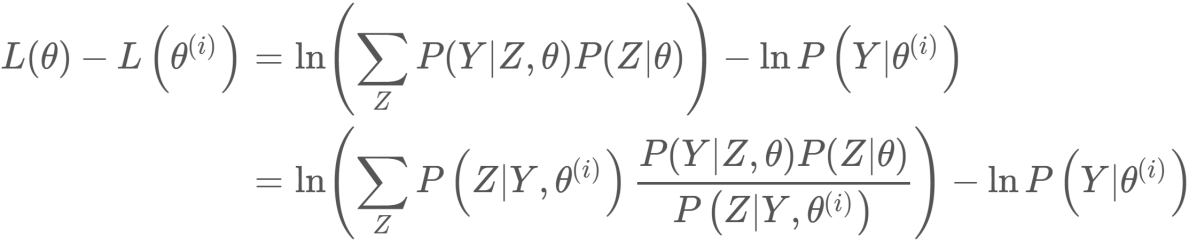
其中，是随机变量，是凸函数，表示的期望。

**EM算法的推导**

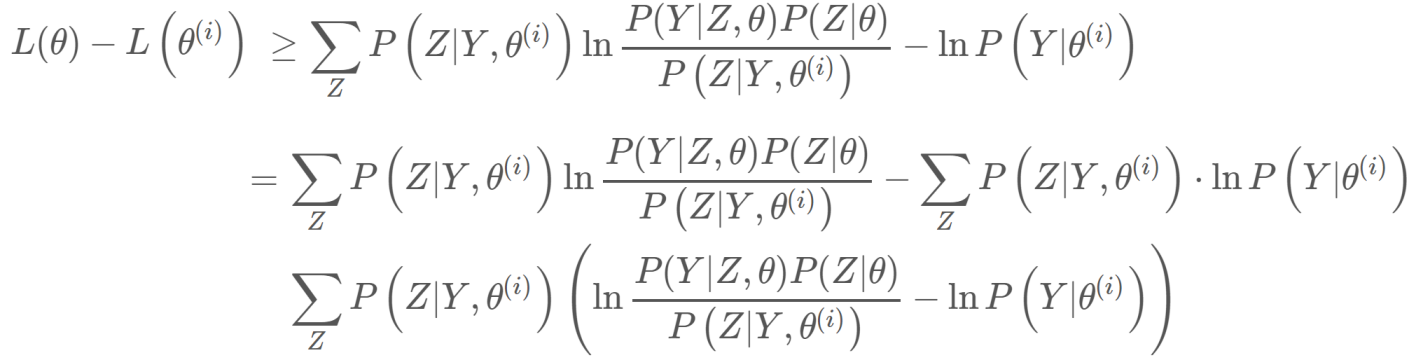
我们面对一个含有隐变量的概率模型，目标是极大化观测数据Y关于参数的对数似然函数，即极大化：

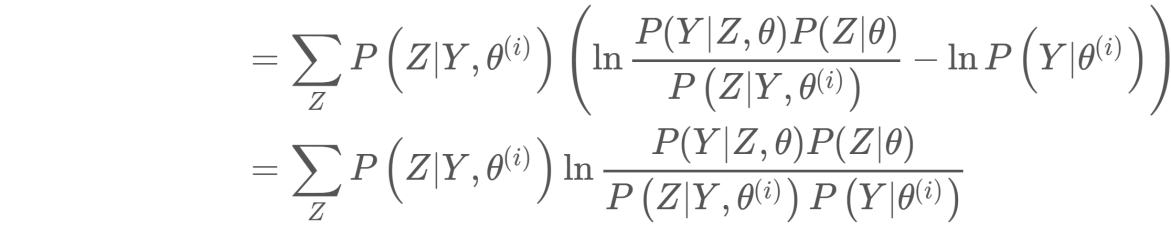


注意到这一极大化的主要困难是上式中有未观测数据Z，并且包含和（Z为连续型时为积分）的对数。EM算法采用的是通过迭代逐步极大化：假设在第次迭代后的估计值是，我们希望新的估计值能使增加，即，并逐步达到极大值。为此，我们考虑两者的差：

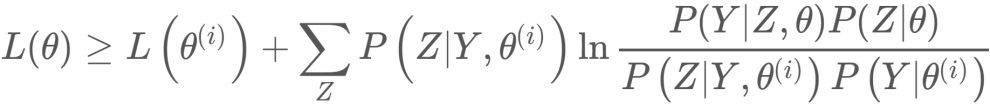


套用琴生不等式可得：

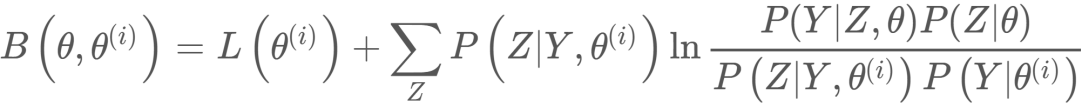




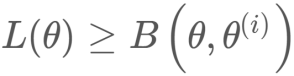
则



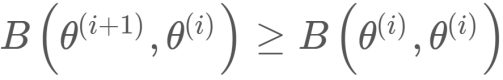
令



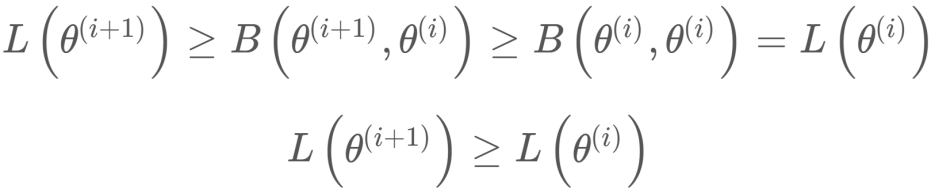
则



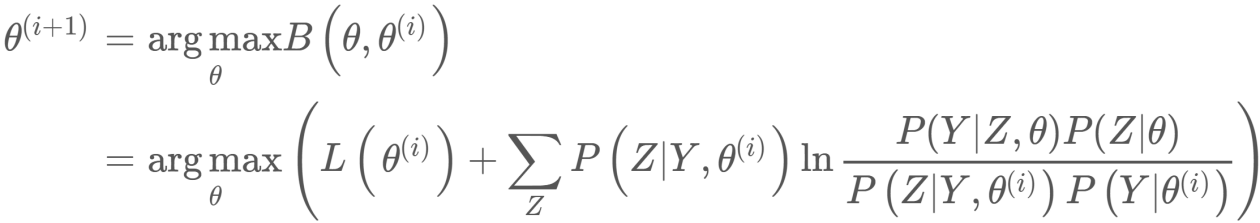
即函数是的一个下界，此时若设使得达到极大，即



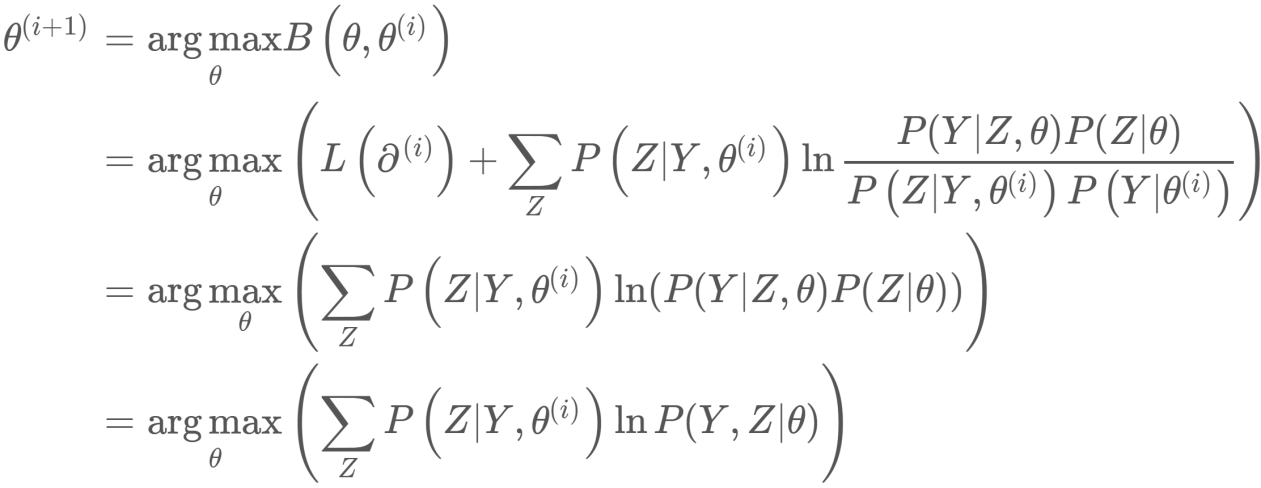
由于，所以可以进一步推得：

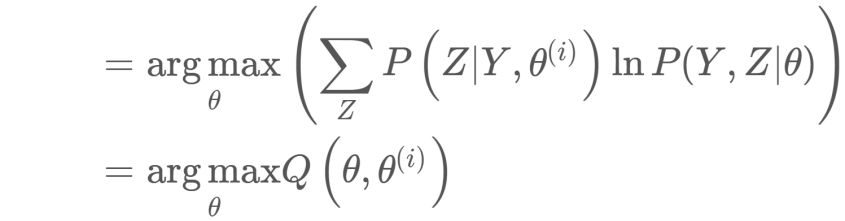


因此，任何可以使增大的，也可以使增大，于是问题就转化为了求解能使得达到极大的，即



这里需要极大化，所以可以省略与无关的项，即





到此即完成了EM算法的一次迭代，求出的作为下一次迭代的初始。综上可以总结出EM算法的“E步”和“M步”分别为：  
E步：计算完全数据的对数似然函数关于在给定观测数据Y和当前参数下对未观测数据Z的条件概率分布的期望，即函数：



M步：求使得函数到达极大的。

将函数展开得



展开 

化解后面的ln 

后面第一项求和中单独拿出第一项



乘法展开



展开最前面的求和符号



这里，第二项也可以类似于第一项这样展开，得

  （1）

这里，只要计算出一个项，其他也都能计算，则以第一项为例



前面乘积中单独拿出第一个因子



再展开前面的求和符号，对展开，即将的每个取值都写出来





提取公因式得





两项的求和项是一样的，再次提取求和项得

前面括号中可以将写为求和形式为

 （2）

这里，后面部分的值为1，单独考察这项



先对单独展开



再对单独展开



提取公因式



提取求和项



再将写为求和形式



同理，可以依次将拆开，为



可以看到，这里每一项的值为1，所以

 （3）

（3）代入（2）有

 （4）

将（4）再带回函数（1）得





写为求和形式为

 （5）

再将展开，则函数可写为

（6）

接下来，讨论上式在三硬币模型的每项含义

首先，来看

表示抛掷硬币A为正面时，观测到的正反面的概率，这里有两种取值为1和0，则

表示抛掷硬币A为正面，然后抛掷硬币B也为正面的概率，所以



表示抛掷硬币A为正面，然后抛掷硬币B为反面的概率，所以



所以，可以将写为

 （7）

同理，可以将写为

 （8）

然后，来看

表示在观测值时硬币A出现正面的概率，求此概率，可以变形为

 （9）

为何要化为（9）的形式？

可以看到，（9）的分子即为（7），而分母在最初时即已得到，即

 （10）

所以将（7）与（10）代入（9）中得



上式即为李航《统计学习方法》的式9.5



则

 （11）

将（7）、（8）、（11）代入（6）的函数得

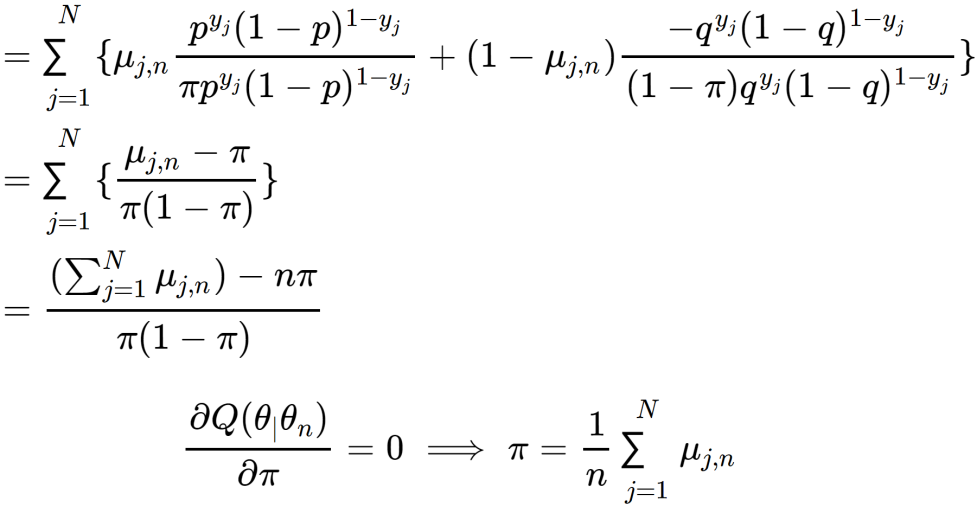
 （12）

接下来，求使得函数极大化的，即M步

这里，只需分别对函数中的求偏导，并令其等于0即可。

首先，对求导得



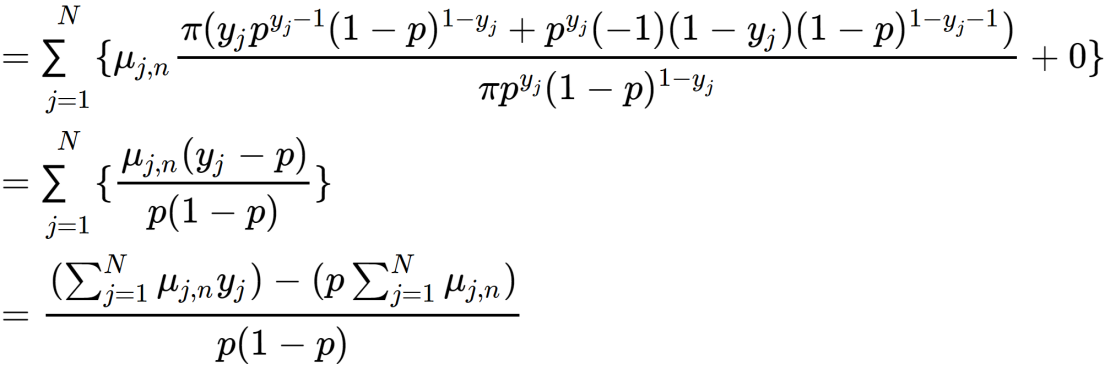


令可解得，令这个为得



即《统计学习方法》的式9.6，同样，对求导得





令可解得



即《统计学习方法》的式9.7，再对求导令其等于0得



即《统计学习方法》的式9.8

利用以上的迭代公式，即可求得最终的参数，这里，假设模型参数的初值为

，，

根据公式9.5，对与均有

利用迭代公式9.6、9.7、9.8得

，，

再根据公式9.5，得

继续迭代，得

，，

第二次迭代时参数就不再变化了，因此迭代停止，于是模型参数θ的极大似然估计就的出来了，即

，，

如果取初值，，，那么得到的模型参数的极大似然估计是，，，这就是说，EM算法与初值的选择有关，选择不同的初始值可能得到不同的参数估计值。